technologies

Article

Criterion Validity of iOS and Android Applications to Measure Steps and Distance in Adults

Manolis Adamakis

School of Education, University College Cork, T12KX72 Cork, Ireland; manosadam@phed.uoa.gr

Abstract: The growing popularity of physical activity (PA) applications (apps) in recent years and the vast amounts of data that they generate present attractive possibilities for surveillance. However, measurement accuracy is indispensable when tracking PA variables to provide meaningful measures of PA. The purpose of this study was to examine the steps and distance criterion validity of freeware accelerometer-based PA smartphone apps, during incremental-intensity treadmill walking and jogging. Thirty healthy adults (25.9 ± 5.7 years) participated in this cross-sectional study. They were fitted with two smartphones (one with Android and one with iOS operating systems), each one simultaneously running four different apps (i.e., Runtastic Pedometer, Accupedo, Pacer, and Argus). They walked and jogged for 5 min at each of the predefined speeds of 4.8, 6.0, and 8.4 km/h on a treadmill, and two researchers counted every step taken during trials with a digital tally counter. Validity was evaluated by comparing each app with the criterion measure using repeated-measures analysis of variance (ANOVA), mean absolute percentage errors (MAPEs), and Bland–Altman plots. For step count, Android apps performed slightly more accurately that iOS apps; nevertheless, MAPEs were generally low for all apps (<5%) and accuracy increased at higher speeds. On the other hand, errors were significantly higher for distance estimation (>10%). The findings suggest that accelerometer-based apps are accurate tools for counting steps during treadmill walking and jogging and could be considered suitable for use as an outcome measure within a clinical trial. However, none of the examined apps was suitable for measuring distance.

Keywords: accelerometer; accuracy; step count; physical activity measurement; software; operating system; pedometer; walking; jogging; monitoring

check for updates

Citation: Adamakis, M. Criterion Validity of iOS and Android Applications to Measure Steps and Distance in Adults. *Technologies* **2021**, *9*, 55. https://doi.org/10.3390/technologies9030055

Academic Editors: James W. Navalta and Alessandro Tognetti

Received: 31 May 2021
Accepted: 26 July 2021
Published: 29 July 2021

1. Introduction

The World Health Organisation has recently updated the guidelines on physical activity (PA) and sedentary behaviour [1]. New guidelines for adults now specify a target range of 150 to 300 min of moderate-intensity and 75 to 150 min of vigorous-intensity PA, compared with the previous guidelines that focused on achieving at least 150 min of moderate-intensity or 75 min of vigorous-intensity activity per week [2]. Updated PA guidelines pose many challenges to PA surveillance; however device-based measures may facilitate surveillance [3].

The growing popularity of wearable PA monitors and fitness applications (apps) in recent years and the vast amounts of data that they generate present attractive possibilities for surveillance [4]. This approach of PA measurement has been currently used in various settings, before [5] and during the COVID-19 pandemic [6–9]. For example, PA data from 717,527 people across 111 countries, in which step count was measured using smartphones with in-built accelerometry, showed that the average user walked 4961 steps per day before the pandemic [5]. Moreover, data from 55,404 users from 187 countries revealed that within 10 days of the pandemic declaration, there was a 5.5% decrease in mean steps, and within 30 days, there was a 27.3% decrease in mean steps [9].

Unfortunately, these consumer-targeted wearable technologies intended for general wellness purposes are not required to undergo a standardised and transparent evaluation process for ensuring their quality and accuracy, and usually product manufacturers present only the minimum requirements when releasing information [10]. The consequence of an app lacking evidence-based support depends on its intended use. For example, to provide meaningful PA estimates, the accurate and reliable assessment derived from wearable monitors and apps is necessary, for any research study where PA is either an outcome measure or an intervention. As mentioned by Nelson et al. [11], measurement accuracy is indispensable when tracking PA variables to provide meaningful measures of PA. The automation of human health monitoring can be problematic even with evidence-based apps, but it makes consumers especially vulnerable if there are no scientific data supporting the apps' claims [12].

A recent review of treadmill-based validation of step counting wearable technologies estimated that mean absolute percentage error (MAPE) values were 7% to 11% for wrist-worn, 1% to 4% for waist-worn, and ≤1% for thigh-worn monitors [13]. Two other systematic reviews examined the validity of Fitbit and Garmin monitors. Fitbit devices were likely to meet acceptable accuracy for step count approximately half the time, with a tendency to underestimate steps in controlled testing and overestimate steps in free-living settings [14]. Garmin activity trackers were also accurate in step counting, with acceptable MAPE values, however distance validity was generally low, with MAPE exceeding the acceptable limits [15].

The validation of PA apps is even more urgent nowadays since apps with low accuracy and insufficient selection justification [16,17] are used in large observational studies with big datasets. Two of the previously mentioned studies [5,9] used the Argus Azumio app to capture large-scale PA data (i.e., steps). However, Brodie et al.'s [17] research on Argus validity revealed significant undercounting (15–66%) by iOS smartphones and extraordinarily large error ranges (0–200% of steps taken) for both Android and iOS smartphones. The authors further suggested that there might be an improved app accuracy for walks of at least 10 steps for apps installed in newer Android and iPhone models (i.e., Android 7.1 compared to Android 8.1).

Regarding the overall validity, a recent review concluded that there is conflicting and insufficient evidence on the validity and reliability of apps for measuring PA. Nevertheless, speed and place where the smartphone is carried seem to have an impact on validity, as absolute errors decreased with higher speeds [18]. For example, Leong and Wong [19] tested three different apps (Runtastic Pedometer, Pedometer Pacer Works, and Pedometer Tayatau) carried at three different places (hip level, waist, and upper arm) and for five different speeds. Steps' mean percentage difference increased as speed decreased and, in general, it was higher than 10% for both the Runtastic and Pedometer Pacer Works, except for Runtastic at higher speeds when carried at the hip level and waist. Furthermore, Orr et al. [20] used eight different speeds (four for walking and four for running) and three different apps (Accupedo, Moves, and Runtastic Pedometer) running in a smartphone carried in the hand. Mean percentage difference was lower than 10% for Accupedo and Runtastic only during running when compared against manual and self-counting. Höchsmann et al. [21] assessed six accelerometer apps (Apple Health, Samsung S Health, Moves, Runtastic Pedometer, Accupedo, and Pacer) running in two different smartphones (iPhone SE and Samsung Galaxy S6 Edge) at four speeds. In this study, all smartphone apps showed high accuracy and low variability for all treadmill conditions, independent of the phone's position. In general, variability decreased, and accuracy increased with increasing walking speed.

On the contrary, a study conducted by Konharn et al. [22] in young adults examined the validity of three apps (Runtastic Pedometer, Footsteps Pedometer, and Walker Pedometer) and found that the apps were not accurate in counting most of the measured variables (e.g., steps and distance) and data fell significantly lower in these parameters than those measured with standard-reference instruments. Xie et al. [23] tested two apps (Dongdong and Ledongli) running in a smartphone placed in the pocket for two laps of 400 m and reported a mean percentage difference lower than 10% for steps and higher than 10% for distance. A final research that evaluated the accuracy of five apps (Moves, Google-Fit, Runtastic Pedometer, Accupedo, and Samsung S Health) while wearing four Samsung Galaxy S4 smartphones on various body locations, concluded that Samsung S Health recorded the lowest mean bias for each given location, while Accupedo at the waist location recorded the largest mean step overestimation and Accupedo at the hand location produced the largest step underestimation of all apps and locations [24].

Since there is an apparent potential of PA apps to measure and promote PA [25] and due to the conflicting validity evidence that currently exists, there is a need to carry out more studies of high methodological quality. Thus, the purpose of the present study was to validate step count and distance travelled of eight freeware accelerometer-based apps (four Android and four iOS apps) in a sample of healthy adults. A secondary purpose was to compare the validity of similar apps running simultaneously in two operating systems, i.e., Android and iOS, since none of the previous research approaches has directly compared apps running in different operating systems. Based on the evaluation frameworks proposed by Keadle et al. [26] and Johnston et al. [27], a semi-structured validation study design in laboratory conditions was used, which included laboratory treadmill walking and jogging at three incremental intensities (i.e., 4.8, 6.0, and 8.4 km/h).

2. Materials and Methods

2.1. Study Design

This study utilised a cross-sectional, repeated-measures research design, investigating the differences in recorded steps and distance values for eight PA smartphone apps. Thirty healthy adults, with no contraindications for exercise and no known orthopaedic limitations that would prevent them from completing the assessments, participated. All adults read and signed an informed consent document approved by the Social Research Ethics Committee of University College Cork, informing them of the risks and benefits of the study.

Participants reported to the researchers twice. During the first visit, anthropometric measures were obtained in controlled laboratory settings. For the second visit (i.e., a week after the first visit), participants returned to the laboratory for treadmill-based walking and jogging. They were instructed to wear their own sports shoes and clothing.

The participants were fitted with two smartphones (one Samsung Galaxy S8 and one iPhone 8), each one simultaneously running four different apps. The two smartphones were strapped close to the body on a waist-worn elastic belt over the left hip, near the anterior axillary line, and were counterbalanced for anterior and posterior placement on the hip among participants. All apps were updated with the participants' age, sex, height, dominant hand, weight, and step length, and the apps' software was updated to the latest available version. Smartphones were set to airplane mode to avoid interactions with the mobile phone providers (i.e., no data connection), and all apps were activated simultaneously.

Prior to testing, participants were familiarised with the motorised treadmill, and then they had to perform three treadmill-based tests. They walked or jogged for 5 min at each of the predefined speeds of 4.8 km/h (light intensity), 6.0 km/h (moderate intensity), and 8.4 km/h (vigorous intensity) on a treadmill (Zebris FDM-T, Zebris Medical GmbH, Isny, Germany) with 0° incline. All trials were completed on the same day and in a randomised order, with 5 min of rest between the various conditions, while all apps were paused simultaneously. The 5 min of each treadmill condition included the time the treadmill increased the walking or jogging speed. During pause and between the transition from 6.0 to 8.4 km/h, all apps' specific settings were changed from the walking to the running option. At the end of each trial, initially data were stored manually, and at a later time, were uploaded to the related apps' software.

Distance was objectively recorded by the in-built function of the treadmill, and the results were used as the criterion measure. The criterion measure for steps was two manual counters who objectively measured steps with the use of a hand-held counter device (GOGO Four Digit Hand Tally Counter, atafa.com). For all trials, they observed the leg movement of the participants and were separated so they could not view each other's thumb motion nor hear the "clicking" from the counter device. This prevented any synchronised counting between the two. The reliability of this method was tested by comparing a video recording for two walking and running video sequences of two participants. An intra-class correlation coefficient value of 0.99 was obtained through the analyses of the video sequences and the steps recorded by the researchers.

2.2. Participants

A power calculation with findings of observed step counts (correlation of 0.50), alpha two-tailed value of 0.05, and a power of 0.80 indicated a sample size of 29 participants. In total, 30 healthy adults (n = 11 males, n = 19 females) with an age range of 19–43 years (25.9 ± 5.7 years), body mass index range of 17.8–30.5 kg/m^2 (24.4 ± 3.9 kg/m^2) were screened and participated in the study (with no dropouts).

2.3. Antropometric Assessment

Standing height was measured to the nearest 0.1 cm using a wall-mounted Harpenden stadiometer (Harpenden, London, UK) using standard procedures. Body mass was measured with participants in light clothes and bare feet on an electronic scale (Omron BF-511) to the nearest 0.1 kg. Body mass index was calculated as weight (kg)/height squared (m^2).

Regarding step length estimation, the operational definition of a step for specific use in treadmill-based device validation purposes was used [13]: "a foot strike following the complete lifting of that foot from the surface of the treadmill belt" (p. 847). The average walking step length was calculated by performing 20 normal steps and measuring the distance between the start and end line, then dividing the total distance by 20 steps. The same procedure was followed to calculate jogging step length. All anthropometric measurement results are presented in Table 1.

Table 1. Participants' characteristics (Mean ± SD).

	Males (n = 11)	**Females (n = 19)**	**Total (n = 30)**
	M ± SD	M ± SD	M ± SD
Age (years)	26.0 ± 6.6	25.8 ± 5.2	25.9 ± 5.7
Weight (kg)	83.2 ± 16.2	63.4 ± 9.3	70.7 ± 1.7
Height (m)	1.79 ± 0.09	1.64 ± 0.06	1.70 ± 0.10
Body mass index (kg/m^2)	24.7 ± 3.7	23.6 ± 3.9	24.4 ± 3.9
Resting heart rate (bpm)	72.5 ± 7.3	68.5 ± 5.5	70.0 ± 6.4
Walking step length (cm)	74.5 ± 6.4	62.2 ± 6.2	66.7 ± 8.6
Running step length (cm)	101.5 ± 7.5	80.9 ± 5.9	88.5 ± 11.9

2.4. Accelerometer-Based Apps

This study used one Samsung Galaxy S8, based on the Android 10.1 operating system, and one iPhone 8, based on iOS 12.1 operating system. Inclusion criteria for all apps were retrieved from previous protocols [28,29]: (1) free of charge indefinitely after download, applications with a free trial period of finite length were excluded; (2) full and efficient functionality after downloading, without additional software download being necessary; (3) functionality only through the built-in accelerometer (no GPS or 4G/5G signal); (4) ability to record the number of steps taken, average speed, total distance, and energy expenditure; (5) manual input of demographic and anthropometric data (sex, age, weight, height, and step length for walking and running); (6) manual choice of activity type (i.e., walking or running); (7) among the most popular and downloadable applications, according to users' ratings and number of downloads from the Google Play Store and App Store; (8) available for use in both Android and iOS smartphones.

Based on the previously described criteria, four accelerometer-based apps were selected, which were installed in Android and iOS smartphones: Runtastic Pedometer (Runtastic GmbH/Adidas, Pasching near Linz, Austria), Accupedo (Corusen LLC, Keller, TX, USA), Pacer (Pacer Health Inc., Miami Lakes, FL, USA), and Argus (Azumio Inc., Redwood City, CA, USA).

2.5. Statistical Analysis

Descriptive analyses were conducted to examine associations with the criterion measures. Six separate repeated-measures analysis of variance (RM-ANOVA) statistical tests were performed to assess differences between all apps and criterion measures for distance and step count at 4.8, 6.0, and 8.4 km/h (three RM-ANOVA tests for steps and three for distance, respectively). When the main RM-ANOVA test statistic was significant (i.e., criterion versus all eight apps), post hoc pairwise comparisons with Bonferroni correction were performed to determine where significant differences existed. The significance level was set at $p < 0.05$ and the partial η^2 was presented as a measure of effect size for F-tests. A partial η^2 value between 0.01 and 0.06 was associated with a small effect, between 0.06 and 0.14 with a medium effect, and 0.14 or greater with a large effect [30].

To more thoroughly investigate the apps' validity, the use of MAPE values and Bland–Altman limits of agreement analysis was implemented [27]. MAPE values were also calculated to provide an indicator of overall measurement error (MAPE = ([monitor measurement-criterion measure]/criterion measure) × 100) and was used as an outcome measure. A smaller MAPE represents better accuracy. Johnston et al. [27] recommend MAPE $\leq$ 5%, if the PA monitor is to be used as an outcome measure within a clinical trial or as an alternative gold-standard measurement tool for step counting, and MAPE $\leq 10-15$% if the device is being validated for use by the general population.

To further evaluate individual variations in a more systematic way, Bland–Altman plots with corresponding 95% limits of agreement and fitted lines (from regression analyses between mean and difference) with their corresponding parameters (i.e., intercept and slope) were presented [31,32]. Dashed lines represent the 95% prediction interval, and solid lines represent the mean errors. A fitted line that provides a slope of 0 and an intercept of 0 exemplifies perfect agreement, while a statistically significant slope suggests that there is proportional systematic bias (i.e., the app gives values that are higher or lower than those from the criterion by an amount that is proportional to the level of the measured variable). The statistical analyses were performed with SPSS version 23.0 for Windows (IBM SPSS Corp., Armonk, NY, USA) and MedCalc 12.7 (MedCalc Software bvba).

3. Results

3.1. Step Count

Participants averaged 569 ± 39 steps when walking at 4.8 km/h, 622 ± 36 steps when walking at 6.0 km/h, and 751 ± 340 steps when jogging at 8.4 km/h, respectively. The RM-ANOVA for walking at 4.8 km/h ($F(8,232) = 1.94$, $p = 0.045$, $\eta^2 = 0.06$), 6.0 km/h ($F(8,232) = 3.64$, $p = 0.001$, $\eta^2 = 0.11$), and jogging at 8.4 km/h ($F(8,232) = 6.88$, $p < 0.001$, $\eta^2 = 0.19$) were statistically significant, with various effect sizes. The post hoc pairwise comparisons with Bonferroni corrections showed that all iOS apps, except for Argus iOS at 4.8 km/h ($F(1,29) = 3.62$, $p = 0.067$, $\eta^2 = 0.11$) and 6.0 km/h ($F(1,29) = 2.69$, $p = 0.112$, $\eta^2 = 0.09$), differed statistically significantly from the criterion during all conditions ($p < 0.05$). All Android apps showed similar results with the criterion ($p > 0.05$) (Table 2).

During walking at 4.8 km/h, the MAPE was low for all apps, ranging from 1.51% (Accupedo Android and Argus Android) to 3.20% for Argus iOS. During walking at 6.0 km/h and jogging at 8.4 km/h, the MAPE was even lower for all Android apps (0.41–0.82%) and higher for iOS apps (e.g., 2.75% for Argus iOS at 6.0 km/h). Overall, all the MAPE values were low (Figure 1).

Table 2. Results of repeated measures ANOVA for *step count* and comparison with criterion measure.

	4.8 km/h				6.0 km/h				8.4 km/h			
	M (SD)	F	p	95% CI	M (SD)	F	p	95% CI	M (SD)	F	p	95% CI
Criterion	569 (39)	-	-	-	622 (36)	-	-	-	751 (40)	-	-	-
Accupedo iOS	562 (56)	12.87	0.001	0.11–15.02	616 (34)	20.21	<0.001	1.35–11.25	745 (40)	9.65	0.004	(−0.81)–12.54
Accupedo Android	566 (42)	1.37	0.251	(−7.07)–14.07	625 (40)	1.75	0.196	(−9.43)–4.30	752 (39)	1.16	0.039	(−3.28)–1.75
Pacer iOS	562 (56)	12.87	0.001	0.11–15.02	616 (34)	20.21	<0.001	1.35–11.25	745 (40)	9.65	0.004	(−0.81)–12.54
Pacer Android	565 (42)	1.55	0.224	(−6.89)–14.35	625 (40)	1.75	0.196	(−9.43)–4.30	752 (39)	1.16	0.039	(−3.28)–1.75
Runtastic iOS	558 (41)	6.22	0.019	(−4.55)–26.35	616 (34)	20.21	<0.001	1.35–11.25	743 (41)	7.05	0.013	(−2.83)–19.90
Runtastic Android	573 (54)	0.25	0.624	(−32.23)–24.30	625 (40)	1.75	0.196	(−9.43)–4.30	752 (39)	1.16	0.039	(−3.28)–1.75
Argus iOS	555 (52)	3.62	0.067	(−12.22)–40.69	607 (56)	2.69	0.112	(−17.70)–48.30	745 (40)	9.65	0.004	(−0.81)–12.54
Argus Android	566 (42)	1.37	0.251	(−7.07)–14.07	625 (40)	1.75	0.196	(−9.43)–4.30	752 (39)	1.16	0.039	(−3.28)–1.75

Note. CI: confidence interval; degrees of freedom: (1,29) for all post hoc tests.

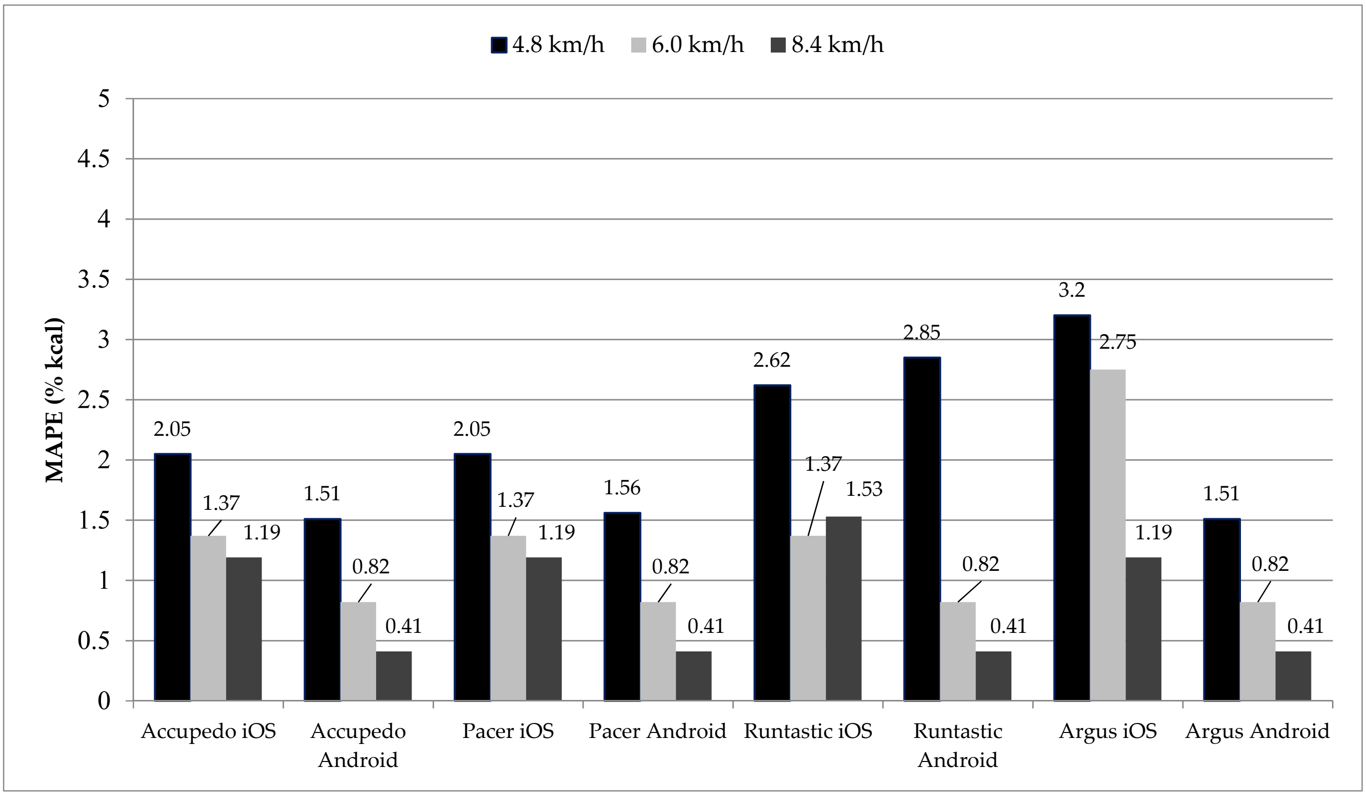

Figure 1. MAPE (% steps) of PA monitors and apps compared with criterion measure.

Bland–Altman results for step count for the three conditions are presented in Table 3, and all Bland–Altman plots are included in the Supplementary file (Figures S1–S48). For walking at 4.8 km/h, the plots revealed the narrowest 95% limits of agreement for Accupedo Android and Argus Android (difference = 3.50 steps), while values were the widest for Runtastic iOS (difference = 10.90 steps) and Argus iOS (difference = 14.23 steps). During walking at 6.0 km/h and jogging at 8.4 km/h, the narrowest 95% limits of agreement were observed for all Android apps (difference = −2.57 and −0.77 steps, respectively), and values were wider for all iOS apps. The widest values were observed for Argus iOS at 6.0 km/h (difference = 15.30 steps) and Runtastic iOS at 8.4 km/h (difference = 8.53 steps). The slopes for the fitted lines were not statistically significant ($p < 0.05$), suggesting that there were no significant patterns of proportional systematic steps' underestimation or overestimation for these apps.

Table 3. *Step count* Bland–Altman results at various speeds.

	4.8 km/h					6.0 km/h					8.4 km/h				
App	M diff	95% CI	Slope	p	95% CI	M diff	95% CI	Slope	p	95% CI	M diff	95% CI	Slope	p	95% CI
Accupedo iOS	7.57	3.25–11.88	0.10	0.063	(−0.01)–0.21	6.30	3.43–9.17	0.08	0.052	0.00–0.15	5.87	2.00–9.73	0.01	0.789	(−0.09)–0.11
Accupedo Android	3.50	(−2.62)–9.62	0.01	0.951	(−0.17)–0.16	−2.57	(−6.54)–1.40	−0.07	0.202	(−0.18)–0.04	−0.77	(−2.22)–0.69	0.01	0.498	(−0.03)–0.05
Pacer iOS	7.57	3.25–11.88	0.10	0.063	(−0.01)–0.21	6.30	3.43–9.17	0.08	0.052	0.00–0.15	5.87	2.00–9.73	0.01	0.789	(−0.09)–0.11
Pacer Android	3.73	(−2.41)–9.88	−0.01	0.895	(−0.18)–0.15	−2.57	(−6.54)–1.40	−0.07	0.202	(−0.18)–0.04	−0.77	(−2.22)–0.69	0.01	0.498	(−0.03)–0.05
Runtastic iOS	10.90	1.96–19.84	0.12	0.286	(−0.11)–0.36	6.30	3.43–9.17	0.08	0.052	0.00–0.15	8.53	1.96–15.11	0.06	0.446	(−0.11)–0.23
Runtastic Android	−3.97	(−20.32)–12.39	0.18	0.407	(−0.26)–0.61	−2.57	(−6.54)–1.40	−0.07	0.202	(−0.18)–0.04	−0.77	(−2.22)–0.69	0.01	0.498	(−0.03)–0.05
Argus iOS	14.23	(−1.07)–29.54	0.15	0.465	(−0.26)–0.55	15.30	(−3.79)–34.39	0.29	0.276	(−0.25)–0.83	5.87	2.00–9.73	0.01	0.789	(−0.09)–0.11
Argus Android	3.50	(−2.62)–9.62	−0.01	0.951	(−0.17)–0.16	−2.57	(−6.54)–1.40	−0.07	0.202	(−0.18)–0.04	−0.77	(−2.22)–0.69	0.01	0.498	(−0.03)–0.05

Note. CI: confidence interval; M diff: Mean difference.

3.2. *Distance*

Participants averaged 0.39 ± 0.01 km during walking at 4.8 km/h, 0.49 ± 0.01 km during walking at 6.0 km/h, and 0.69 ± 0.01 km during jogging at 8.4 km/h, respectively. The RM-ANOVA for walking at 4.8 km/h (F(8,232) = 8.17, $p < 0.001$, $\eta^2 = 0.22$), 6.0 km/h (F(8,232) = 14.55, $p < 0.001$, $\eta^2 = 0.33$), and jogging at 8.4 km/h (F(8,232) = 8.19, $p < 0.001$, $\eta^2 = 0.22$) were statistically significant, with large effect sizes. The post hoc pairwise comparisons with Bonferroni corrections showed that most apps differed statistically significantly from the criterion at 6.0 km/h and 8.4 km/h ($p < 0.05$). At 4.8 km/h, only Argus iOS (F(1,29) = 7.98, p = 0.008, $\eta^2 = 0.22$), Argus Android (F(1,29) = 11.04, p = 0.002, $\eta^2 = 0.28$) and Runtastic Android (F(1.29) = 10.45, p = 0.003, $\eta^2 = 0.27$) differed statistically significantly from the criterion. All comparisons with the criterion are presented in Table 4.

Table 4. Results of repeated measures ANOVA for *distance* (km) and comparison with criterion measure.

	4.8 km/h				6.0 km/h				8.4 km/h			
	M (SD)	F	p	95% CI	M (SD)	F	p	95% CI	M (SD)	F	p	95% CI
Criterion	0.39 (0.01)	-	-	-	0.49 (0.01)	-	-	-	0.69 (0.01)	-	-	-
Accupedo iOS	0.41 (0.05)	1.67	0.207	(−0.05)–0.02	0.46 (0.05)	7.80	0.009	(−0.01)–0.07	0.63 (0.08)	20.93	<0.001	0.01–0.11
Accupedo Android	0.40 (0.07)	0.16	0.696	(−0.05)–0.04	0.45 (0.04)	24.57	<0.001	0.01–0.07	0.60 (0.07)	51.03	<0.001	0.05–0.14
Pacer iOS	0.39 (0.04)	0.73	0.401	(−0.02)–0.03	0.45 (0.06)	11.99	0.002	0.00–0.08	0.63 (0.12)	7.39	0.011	(−0.02)–0.13
Pacer Android	0.38 (0.08)	1.25	0.274	(−0.03)–0.06	0.43 (0.06)	31.53	<0.001	0.02–0.11	0.63 (0.13)	6.22	0.019	(−0.03)–0.15
Runtastic iOS	0.39 (0.10)	0.01	0.923	(−0.06)–0.06	0.53 (0.14)	1.65	0.209	(−0.13)–0.06	0.78 (0.17)	7.56	0.010	(−0.20)–0.03
Runtastic Android	0.49 (0.16)	10.45	0.003	(−0.20)–0.01	0.54 (0.09)	6.76	0.15	(−0.10)–0.02	0.73 (0.12)	3.75	0.063	(−0.12)–0.03
Argus iOS	0.44 (0.09)	7.98	0.008	(−0.10)–0.01	0.53 (0.13)	2.44	0.129	(−0.12)–0.05	0.68 (0.17)	0.19	0.665	(−0.10)–0.12
Argus Android	0.64 (0.41)	11.04	0.002	(−0.52)–0.02	0.65 (0.18)	24.67	<0.001	(−0.27)–(−0.05)	0.77 (0.20)	4.74	0.038	(−0.21)–0.05

Note. CI: confidence interval; degrees of freedom: (1,29) for all post hoc tests.

Figure 2 reports the MAPE for all iOS and Android apps. During walking at 4.8 km/h, the magnitude of errors was lower for Pacer iOS (7.46%) and Accupedo iOS (9.69%), while error rates for all other were above 12.00% (12.26–35.15%). During walking at 6.0 km/h, the magnitude of errors was lower for Accupedo Android (9.78%), followed by Accupedo iOS (10.41%). Error rates for the remaining apps ranged from 12.45% to 36.58%. Lastly, all MAPE values for jogging at 8.4 km/h were higher than 10.0%, ranging from 12.18% for Accupedo iOS to 21.01% for Argus iOS and Argus Android.

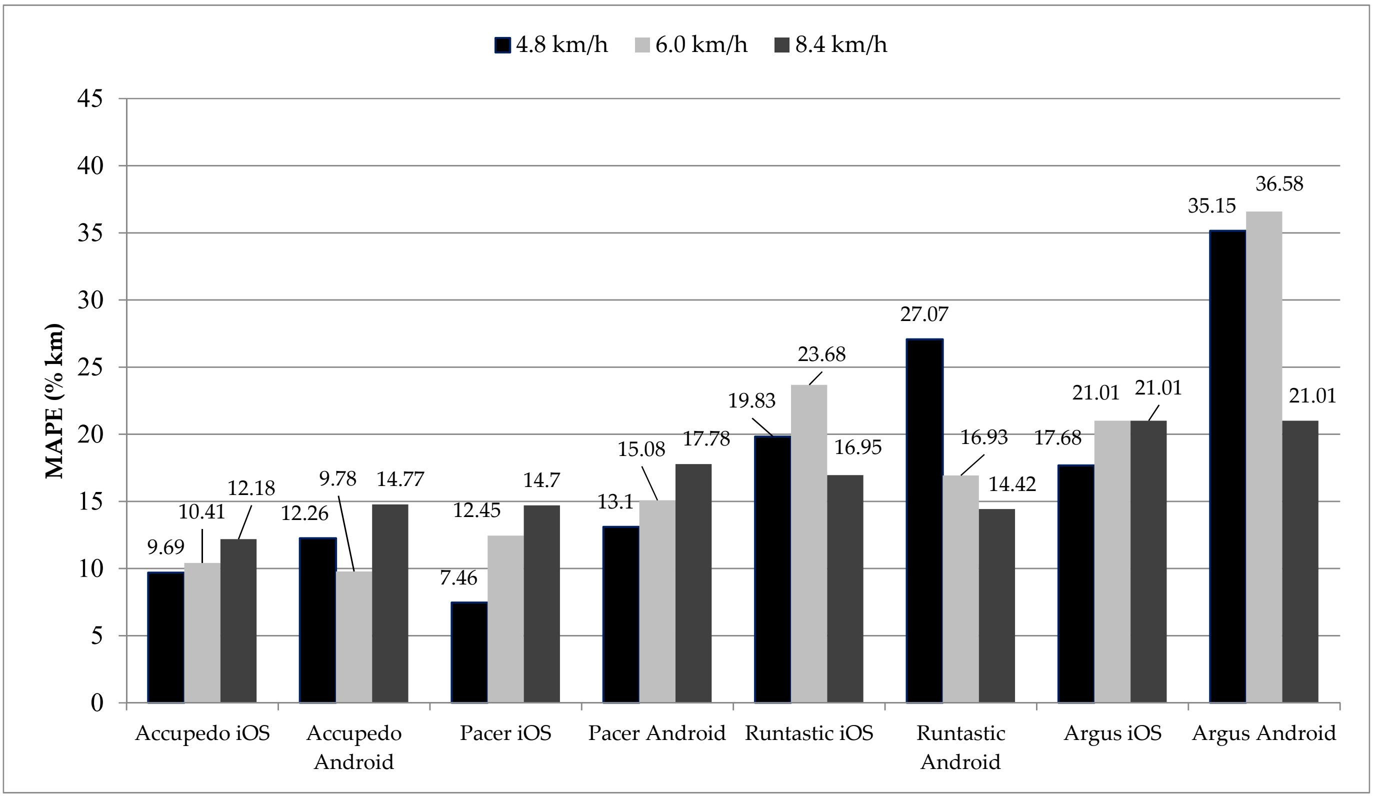

Figure 2. MAPE (% km) of PA monitors and apps compared with criterion measure.

Bland–Altman results for distance for the three conditions are presented in Table 5, and all Bland–Altman plots are included in Figures S1–S48. For walking at 4.8 km/h, the plots revealed the narrowest 95% limits of agreement for Runtastic iOS (difference = 0.00 km) and Sports Tracker (difference = −0.01 km), and slightly wider values for Accupedo (difference = −0.02 km) and Accupedo iOS, Accupedo Android, and Pacer iOS (difference = 0.01 km), while values were the widest for Runtastic Android (difference = −0.10 km) and Argus Android (difference = −0.25 km). During walking at 6.0 km/h and jogging at 8.4 km/h, the narrowest 95% limits of agreement were observed for Accupedo iOS and Runtastic iOS at 6.0 km/h (difference = 0.03 km) and for Argus iOS (difference = 0.01 km) and Runtastic Android at 8.4 km/h (difference = −0.04 km). The widest values were observed for Argus Android at 6.0 km/h (difference = −0.16 km) and Accupedo Android at 8.4 km/h (difference = 0.10 km). In general, the slopes for the fitted lines were not statistically significant ($p < 0.05$), suggesting that there were no significant patterns of proportional systematic bias for these apps, except for Accupedo iOS during walking at both speeds ($p = 0.029$ and $p < 0.001$, respectively), Argus iOS at 4.8 km/h ($p < 0.001$), and Accupedo Android at 6.0 km/h ($p = 0.015$).

Table 5. Distance Bland–Altman results at various speeds.

	4.8 km/h					6.0 km/h					8.4 km/h				
App	M diff	95% CI	Slope	p	95% CI	M diff	95% CI	Slope	p	95% CI	M diff	95% CI	Slope	p	95% CI
Accupedo iOS	−0.01	(−0.03)–0.01	3.53	0.029	0.38–6.68	0.03	0.01–0.05	1.74	<0.001	0.85–2.63	0.06	0.03–0.09	−0.47	0.653	(−2.56)–1.63
Accupedo Android	−0.01	(−0.03)–0.02	−0.75	0.731	(−5.17)–3.67	0.04	0.02–0.06	0.94	0.015	0.20–1.68	0.10	0.07–0.12	0.48	0.644	(−1.63)–2.59
Pacer iOS	0.01	(−0.01)–0.02	0.84	0.485	(−1.60)–3.28	0.04	0.02–0.06	0.59	0.310	(−0.58)–1.76	0.06	0.01–0.10	−1.34	0.413	(−4.64)–1.96
Pacer Android	0.02	(−0.01)–0.04	−0.50	0.833	(−5.30)–4.30	0.07	0.04–0.09	0.76	0.189	(−0.39)–1.91	0.06	0.01–0.11	1.60	0.400	(2.23)–5.43
Runtastic iOS	0.00	(−0.03)–0.04	−4.09	0.157	(−9.86)–1.67	−0.03	(−0.09)–0.02	1.09	0.401	(−1.53)–3.72	−0.09	(−0.15)–(−0.02)	4.78	0.500	0.01–9.54
Runtastic Android	−0.10	(−0.16)–(−0.04)	4.25	0.410	(−6.15)–14.65	−0.04	(−0.08)–(−0.01)	0.93	0.239	(−0.65)–2.52	−0.04	(−0.08)–0.00	−3.06	0.053	(−6.17)–0.05
Argus iOS	−0.04	(−0.07)–(−0.01)	−8.00	<0.001	(−9.28)–3.72	−0.04	(−0.09)–0.01	0.46	0.701	(−1.96)–2.87	0.01	(−0.05)–0.08	−2.29	0.344	(−7.16)–2.58
Argus Android	−0.25	(−0.40)–(−0.10)	20.16	0.111	(−4.96)–45.27	−0.16	(−0.22)–(−0.09)	−0.27	0.864	(3.48)–2.94	−0.08	(−0.15)–(−0.01)	−1.29	0.650	(−7.07)–4.49

Note. CI: confidence interval; M diff: Mean difference.

4. Discussion

The aim of the present study was to examine the validity of four iOS and four Android PA apps in measuring steps and distance during incremental-intensity treadmill walking and jogging in a sample of healthy adults. To our knowledge, this is the first study to examine and directly compare these estimates between the two major smartphone operating systems.

The primary finding regarding step count was that all freeware accelerometer-based apps were valid in all conditions, with MAPE values well below 5%, and no systematic biases were observed. The lowest error was estimated for all Android apps during jogging at 8.4 km/h, while iOS apps had slightly higher errors compared to the criterion measure. The highest error was observed for Argus iOS during walking. Furthermore, no systematic biases were observed for these apps. Regarding distance validity, it was found that Android and iOS apps were not valid in all conditions, with high individual errors (>10%). The lowest errors were observed for Pacer iOS and Accupedo iOS during walking at 4.8 km/h, while highest errors were estimated for Runtastic and Argus apps.

Some previous validation studies for step count concluded that PA apps were likely to meet acceptable accuracy levels [19–21,33] and accuracy increased at higher speeds [18,21]. The results of the current study supported previous findings. All iOS and Android apps improved their step accuracy at higher speed (i.e., 8.4 km/h), compared to the walking trials. Furthermore, the errors in all conditions were low, resulting in comparable results with studies on wearable PA monitors' validity [13,34,35]. Based on Johnston et al.'s [27] recommendations, these apps installed in Samsung Galaxy S8 and iPhone 8 smartphones have the potential to be used as step outcome measures within clinical trials or as alternative gold-standard measurement tools for step counting. More validation studies should be carried out to further support this outcome.

When comparing iOS and Android apps, Android apps performed slightly more accurately than iOS ones. To our knowledge, no previous studies have directly compared the same apps installed in the two operating systems. One previous study [33] examined the step validity of the Moves app in Samsung Galaxy S4 and iPhone 5s smartphones and found that data from the apps in the two operating systems were only slightly different than observed step counts in two trials (500 step and 1500 steps). In this study, Accupedo, Pacer, Runtastic, and Argus Android apps were more valid than similar iOS apps in all conditions. However, both iOS and Android apps showed high accuracy levels regarding step count with no systematic errors. Furthermore, all apps had lower errors than the ones estimated in previous studies (e.g., Runtastic, Accupedo, and Pacer [19–22,24]; Argus [17]). This significant improvement may be attributed to updated smartphone hardware and software, as well as to improved app step detection algorithms, because in the current research protocol, newer smartphones, operating systems, and apps were included (e.g., iPhone 8 vs. iPhone 5s and SE; Samsung Galaxy S8 vs. Samsung Galaxy S4).

The statement regarding the higher accuracy of newer versions was also supported in Brodie et al.'s [17] study, even though the examined app (i.e., Argus Azumio) was found to have extraordinarily large error ranges for both Android and iOS smartphones. However, similar errors were not estimated in the present study, even though Argus iOS had the highest error between all apps (e.g., 3.20% in 4.8 km/h). Definitely, more high-quality studies should be conducted to further examine the validity of this specific app, mainly because Argus is used in large observational studies with big step count datasets [5,9] and accurate data are more important than low-quality big data in PA monitoring.

A new finding in this study was that most apps had comparable (or even similar) estimates for step count. In previous studies, step differences between the various apps under examination were significant, and these differences were mainly because apps used proprietary methods to detect steps and, hence, differences may exist in the types of movements that are captured as steps, resulting in step count variability [36]. Currently, most apps use the smartphones' built-in step counter sensors, and this approach potentially affects all apps that use these sensors to provide similar step count estimates. Furthermore, the smartphones' position did not impact step detection accuracy, as all smartphones were placed close to the body, around the waist. It is uncertain whether the accuracy would further increase if smartphones were placed in a different position, i.e., around the arm or in the pocket.

Regarding distance validity, all accelerometer-based iOS and Android apps had large individual and group errors for all conditions, usually above 10%, and some apps had systematic errors in various conditions (e.g., Accupedo iOS). Argus apps were the most inaccurate ones, with large MAPE over 20%. These findings are consistent with previous studies, as most PA activity trackers and apps have been found accurate with step counts but lacked accuracy in reporting distance [18,35]. The errors in distance estimation might be a result of inaccurate initial step detection, inappropriate algorithm(s) for the transformation of step count into distance, and/or step length variability during PA. Taking into account that all apps were initially valid in step counting and the same step length was used (which was included in the conversion algorithm), it seems possible that apps used different proprietary distance estimation algorithms, which were not valid. On the contrary, GPS-based apps are more accurate when it comes to distance estimation [37,38], so GPS-based apps should be used instead of accelerometer-based apps when the primary outcome measure is distance.

The main strengths of this study included the criterion-specific selection of freeware accelerometer-based apps available for use in both iOS and Android smartphones and the comparison to criterion measures. Other strengths included a sample consisting of adults, submaximal treadmill walking and jogging trials, and randomisation of these activities to prevent systematic bias in the measurement. Furthermore, the jogging activity was performed at a high speed (i.e., 8.4 km/h), which was not usually selected in previous studies. Lastly, the results were presented according to the selected speeds, so that researchers have direct access to speed-specific results to facilitate future systematic reviews and meta-analyses. Future validation studies may follow a similar methodological approach to select and examine newer PA apps, installed in newer smartphones and operating systems. The results can further guide the consumers in the selection of the most appropriate and valid app(s) to use for capturing steps and distance during PA.

Limitations of this study included the sample size consisting of healthy participants, while future research approaches should include more semi- or un-structured activities in a free-living environment. In addition, future studies should examine the validity of apps during activities of daily living, preferably over a time frame of 2–4 days to assess the suitability of these devices to be used for long-term accelerometry. As mentioned by Brodie et al. [17], to definitively determine the extent that smartphone apps undercount/overcount non-stereotypical gait, walks by larger people, females, and people from different ethnic groups, and much larger heterogeneous samples are required. Finally, the role of the smartphone's optimal position on the human body during exercise should be further investigated.

5. Conclusions

PA tracking monitors and freeware apps have the potential to capture real-time PA data and have been shown to increase daily PA, but the reliability and validity of numerous commercially available apps remain unclear. In this validation study, all accelerometer-based iOS and Android apps returned a high level of step consistency and accuracy during incremental-intensity treadmill walking and jogging and could be considered suitable for use as an outcome measure within a clinical trial. On the other hand, none of these apps was suitable for measuring distance, and GPS-based apps should be used when distance is the primary outcome measure. Since new monitors and apps are released to the consumer market every year, promising improved measurements and user experience, similar high-quality studies should be continuously conducted to generate scientific data supporting the apps' validity.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/technologies9030055/s1, Figures S1–S48: All Bland–Altman plots for step count and distance comparisons with criterion measure (48 plots in total) for the three conditions (i.e., 4.8, 6.0, and 8.4 km/h).

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to reasons of sensitivity and restrictions, as these contain information that could compromise research participants' privacy/consent.

Conflicts of Interest: The author declares no conflict of interest.

References

1. WHO. *Guidelines on Physical Activity and Sedentary Behaviour*; World Health Organization: Geneva, Switzerland, 2020; Licence: CC BY-NC-SA 3.0 IGO; Available online https://www.who.int/publications/i/item/9789240015128 (accessed on 15 May 2021).
2. Bull, F.C.; Al-Ansari, S.S.; Biddle, S.; Borodulin, K.; Buman, M.P.; Cardon, G.; Carty, C.; Chaput, J.P.; Chastin, S.; Chou, R.; et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. *Br. J. Sports Med.* **2020**, *54*, 1451–1462. [CrossRef] [PubMed]
3. Troiano, R.P.; Stamatakis, E.; Bull, F.C. How can global physical activity surveillance adapt to evolving physical activity guidelines? Needs, challenges and future directions. *Br. J. Sports Med.* **2020**, *54*, 1468–1473. [CrossRef]
4. Omura, J.D.; Carlson, S.A.; Paul, P.; Watson, K.B.; Fulton, J.E. National physical activity surveillance: Users of wearable activity monitors as a potential data source. *Prev. Med. Rep.* **2017**, *5*, 124–126. [CrossRef] [PubMed]
5. Althoff, T.; Sosič, R.; Hicks, J.L.; King, A.C.; Delp, S.L.; Leskovec, J. Large-scale physical activity data reveal worldwide activity inequality. *Nature* **2017**, *547*, 336–339. [CrossRef] [PubMed]
6. Ding, D.; Del Pozo Cruz, B.; Green, M.A.; Bauman, A.E. Is the COVID-19 lockdown nudging people to be more active: A big data analysis. *Br. J. Sports Med.* **2020**, *54*, 1183–1184. [CrossRef]
7. McCarthy, H.; Potts, H.; Fisher, A. Physical activity behavior before, during, and after COVID-19 restrictions: Longitudinal smartphone-tracking study of adults in the United Kingdom. *J. Med. Internet Res.* **2021**, *23*, e23701. [CrossRef]
8. Pépin, J.L.; Bruno, R.M.; Yang, R.Y.; Vercamer, V.; Jouhaud, P.; Escourrou, P.; Boutouyrie, P. Wearable activity trackers for monitoring adherence to home confinement during the COVID-19 pandemic worldwide: Data aggregation and analysis. *J. Med. Internet Res.* **2020**, *22*, e19787. [CrossRef]
9. Tison, G.H.; Avram, R.; Kuhar, P.; Abreau, S.; Marcus, G.M.; Pletcher, M.J.; Olgin, J.E. Worldwide effect of COVID-19 on physical activity: A descriptive study. *Ann. Intern. Med.* **2020**, *173*, 767–770. [CrossRef]
10. Bent, B.; Dunn, J.P. Wearables in the SARS-CoV-2 pandemic: What are they good for? *JMIR mHealth uHealth* **2020**, *8*, e25137. [CrossRef]
11. Nelson, M.B.; Kaminsky, L.A.; Dickin, D.C.; Montoye, A.H. Validity of consumer-based physical activity monitors for specific activity types. *Med. Sci. Sports Exerc.* **2016**, *48*, 1619–1628. [CrossRef]
12. Rhea, C.K.; Felsberg, D.T.; Maher, J.P. Toward evidence-based smartphone apps to enhance human health: Adoption of behavior change techniques. *Am. J. Health Educ.* **2018**, *49*, 210–213. [CrossRef]
13. Moore, C.C.; McCullough, A.K.; Aguiar, E.J.; Ducharme, S.W.; Tudor-Locke, C. Toward harmonized treadmill-based validation of step-counting wearable technologies: A scoping review. *J. Phys. Act. Health* **2020**, *17*, 840–852. [CrossRef] [PubMed]
14. Feehan, L.M.; Geldman, J.; Sayre, E.C.; Park, C.; Ezzat, A.M.; Yoo, J.Y.; Hamilton, C.B.; Li, L.C. Accuracy of Fitbit devices: Systematic review and narrative syntheses of quantitative data. *JMIR mHealth uHealth* **2018**, *6*, e10527. [CrossRef]
15. Evenson, K.R.; Spade, C.L. Review of validity and reliability of Garmin activity trackers. *J. Measur. Phys. Behav.* **2020**, *3*, 170–185. [CrossRef] [PubMed]
16. Adamakis, M. Physical activity in the era of mHealth big data: Considerations on accuracy and bias. *SSP J. Sport Sci. Med.* **2019**, *2*, 6–10.
17. Brodie, M.A.; Pliner, E.M.; Ho, A.; Li, K.; Chen, Z.; Gandevia, S.C.; Lord, S.R. Big data vs accurate data in health research: Large-scale physical activity monitoring, smartphones, wearable devices and risk of unconscious bias. *Med. Hypotheses* **2018**, *119*, 32–36. [CrossRef]
18. Silva, A.G.; Simões, P.; Queirós, A.; Rodrigues, M.; Rocha, N.P. Mobile apps to quantify aspects of physical activity: A systematic review on its reliability and validity. *J. Med. Syst.* **2020**, *44*, 51. [CrossRef]
19. Leong, J.Y.; Wong, J.E. Accuracy of three Android-based pedometer applications in laboratory and free-living settings. *J. Sports Sci.* **2017**, *35*, 14–21. [CrossRef]
20. Orr, K.; Howe, H.S.; Omran, J.; Smith, K.A.; Palmateer, T.M.; Ma, A.E.; Faulkner, G. Validity of smartphone pedometer applications. *BMC Res. Notes* **2015**, *8*, 733. [CrossRef] [PubMed]
21. Höchsmann, C.; Knaier, R.; Eymann, J.; Hintermann, J.; Infanger, D.; Schmidt-Trucksäss, A. Validity of activity trackers, smartphones, and phone applications to measure steps in various walking conditions. *Scand. J. Med. Sci. Sports* **2018**, *28*, 1818–1827. [CrossRef]

22. Konharn, K.; Eungpinichpong, W.; Promdee, K.; Sangpara, P.; Nongharnpitak, S.; Malila, W.; Karawa, J. Validity and reliability of smartphone applications for the assessment of walking and running in normal-weight and overweight/obese young adults. *J. Phys. Act. Health* **2016**, *13*, 1333–1340. [CrossRef] [PubMed]
23. Xie, J.; Wen, D.; Liang, L.; Jia, Y.; Gao, L.; Lei, J. Evaluating the validity of current mainstream wearable devices in fitness tracking under various physical activities: Comparative study. *JMIR mHealth uHealth* **2018**, *6*, e94. [CrossRef] [PubMed]
24. Funk, M.D.; Salazar, C.L.; Martinez, M.; Gonzalez, J.; Leyva, P.; Bassett, D.; Karabulut, M. Validity of smartphone applications at measuring steps: Does wear location matter? *J. Measur. Phys. Behav.* **2019**, *2*, 22–27. [CrossRef]
25. Hartung, V.; Sarshar, M.; Karle, V.; Shammas, L.; Rashid, A.; Roullier, P.; Eilers, C.; Mäurer, M.; Flachenecker, P.; Pfeifer, K.; et al. Validity of consumer activity monitors and an algorithm using smartphone data for measuring steps during different activity types. *Int. J. Behav. Nutr. Phys. Act.* **2020**, *17*, 9314. [CrossRef]
26. Keadle, S.K.; Lyden, K.A.; Strath, S.J.; Staudenmayer, J.W.; Freedson, P.S. A framework to evaluate devices that assess physical behavior. *Exerc. Sport Sci. Rev.* **2019**, *47*, 206–214. [CrossRef]
27. Johnston, W.; Judice, P.B.; Molina García, P.; Mühlen, J.M.; Lykke Skovgaard, E.; Stang, J.; Schumann, M.; Cheng, S.; Bloch, W.; Brønd, J.C.; et al. Recommendations for determining the validity of consumer wearable and smartphone step count: Expert statement and checklist of the INTERLIVE network. *Br. J. Sports Med.* **2020**. [CrossRef]
28. Adamakis, M. Preliminary validation study of consumer-level activity monitors and mobile applications for step counting under free living conditions. *J. Mob. Technol. Med.* **2017**, *6*, 26–33. [CrossRef]
29. Adamakis, M. Criterion validity of wearable monitors and smartphone applications to measure physical activity energy expenditure in adolescents. *Sport Sci. Health* **2020**, *16*, 755–763. [CrossRef]
30. Warner, R.M. *Applied Statistics: From Bivariate through Multivariate Techniques*, 2nd ed.; Sage: Los Angeles, CA, USA, 2012.
31. Bland, J.M.; Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *8*, 307–310. [CrossRef]
32. Ludbrook, J. Statistical techniques for comparing measurers and methods of measurement: A critical review. *Clin. Exp. Pharmacol. Physiol.* **2002**, *29*, 527–536. [CrossRef]
33. Case, M.A.; Burwick, H.A.; Volpp, K.G.; Patel, M.S. Accuracy of smartphone applications and wearable devices for tracking physical activity data. *JAMA* **2015**, *313*, 625–626. [CrossRef] [PubMed]
34. Dowd, K.P.; Szeklicki, R.; Minetto, M.A.; Murphy, M.H.; Polito, A.; Ghigo, E.; van der Ploeg, H.; Ekelund, U.; Maciaszek, J.; Stemplewski, R.; et al. A systematic literature review of reviews on techniques for physical activity measurement in adults: A DEDIPAC study. *Int. J. Behav. Nutr. Phys. Act.* **2018**, *15*, 15. [CrossRef] [PubMed]
35. Evenson, K.R.; Goto, M.M.; Furberg, R.D. Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int. J. Behav. Nutr. Phys. Act.* **2015**, *12*, 159. [CrossRef]
36. John, D.; Morton, A.; Arguello, D.; Lyden, K.; Bassett, D. "What Is a step?" Differences in how a step is detected among three popular activity monitors that have impacted physical activity research. *Sensors* **2018**, *18*, 1206. [CrossRef] [PubMed]
37. Adamakis, M. Comparing the validity of a GPS monitor and a smartphone application to measure physical activity. *J. Mob. Technol. Med.* **2017**, *6*, 28–38. [CrossRef]
38. Pobiruchin, M.; Suleder, J.; Zowalla, R.; Wiesner, M. Accuracy and adoption of wearable technology used by active citizens: A marathon event field study. *JMIR mHealth uHealth* **2017**, *5*, e24. [CrossRef]